Calibration standard for determining base proportions of degenerated bases in DNA

The present invention relates to a calibration standard, and furthermore a method that allows to assess and calibrate methods and systems that quantify base compositions at special positions in DNA. The method is characterized by using synthetic, highly reproducible test systems, termed calibration standard. Said test systems are characterized by a) being built by DNA subclone mixtures, and b) providing high numbers of measuring points within one DNA subclone mixture. Referring to a) the method is characterized by the use of mixtures of subclones from one and the same DNA region that show base composition differences at positions of interest. The method is further characterized by taking more than two subclones, that among one another are as unequal as possible and mix them in permutations of different portions. Referring to b) the measuring points cover the range of the measurement method to be assessed or calibrated in an evenly distributed manner. If used to assess a DNA methylation detection method, the test system is able to test the outcome of single steps of said method, and therefore has a huge advantage compared to methods that can only assess the outcome of multiple steps.

It is known from prior art that DNA samples can contain molecule-to-molecule variations which arise in several ways. For example, such single molecule alterations occur spontaneously or represent a precise regulating tool to switch gene expression on or off.

In particular, methylation patterns often show molecule-to-molecule variations. The methylation of DNA is a naturally occurring event that happens in both prokaryotic and eukaryotic organisms. In prokaryotes, DNA methylation provides a way to protect host DNA from digestion by their own restriction enzymes that are designed to eliminate foreign DNA. In higher eukaryotes, DNA methylation acts as another level of gene expression regulation. It has been clearly demonstrated that aberrant methylation is a widespread phenomenon in cancer and may be among the earliest changes during oncogenesis. DNA methylation has also been shown to play a central role in gene imprinting, embryonic development, X chromosome gene silencing and cell cycle

regulation. In many plants and animals, including mammals, DNA methylation consists of the addition of a methyl group to the fifth-carbon position of the cytosine pyrimidine ring via a methyl transferase enzyme. The majority of the DNA methylation in mammals is found in 5'-CpG-3' dinucleotides, but other methylation patterns do exist. In fact, about 80% of all 5'-CpG-3' dinucleotides in mammalian genomes are methylated, and the majority of the 20% that remain unmethylated are found within promoters or in the first exons of genes.

It is obvious that the ability to quantify and detect DNA methylation efficiently and accurately is essential for the study of cancer, gene expression, genetic diseases, and many other important aspects of biology. To this date a number of methods have been developed in order to quantify DNA methylation, such as high performance capillary electrophoresis and methylation-sensitive restriction digestion, which is for example used in a discovery method called 'methylation sensitive arbitrarily primed PCR' (MS-APPCR), but currently the most commonly used technique is a combination of firstly treating the DNA with an agent such as bisulfite or an enzyme such as Activation Induced Deaminase (AID) and in a second step employing common molecular base detection methods. The bisulfite method consists of treating DNA with bisulfite which causes unmethylated cytosines to be converted into uracil while methylated cytosines remain unchanged. The same effect can be achieved when treating DNA with the AID enzyme. The bisulfite modified DNA is usually amplified by PCR and the resulting PCR products are analyzed for example by DNA sequencing or restriction digestion (known as the COBRA method) or most commonly by hybridization techniques such as "Real-time-PCR". The methylation status of the DNA segment is then determined by comparing the sequence information from the bisulfite treated DNA with that of untreated DNA.

Methods for determining base compositions require a high sensitivity and accuracy to quantify the biological fine-tuning of cells in terms of molecule-to-molecule variations. Such methods have to be tested and calibrated by using standards.

Standards consisting of mixtures of DNA molecules which are identical concerning a specific intramolecular feature are known, i.e. a single base type at different positions. For example, a biological standard is prepared by mixing two DNA samples of natural origin. The sequences of these DNA samples have been characterized in advance.

Furthermore, a synthetic standard for analyzing methylation is known. It is prepared by mixing two DNA samples, one of them being completely unmethylated and the other one being completely methylated. Both standards are able to simulate a single base proportion at each position. The level of proportion can be shifted by the mixing ratio of the sample but each mixing ratio results in another calibration standard. Mixtures of e.g. methylated and unmethylated DNA only provide one defined base proportion to be expected after conversion, it is equal at each position. Any problems that might occur from the fact that other positions have other rates are omitted from such system.

In addition, the generation of mixtures of e.g. methylated and unmethylated DNA requires several steps until it can be used to assess a measurement method based on e.g. PCR products of bisulfite treated DNA, such as e.g. a Real-Time-PCR-Assay. Theses steps are firstly the generation of methylated DNA, secondly the conversion of unmethylated cytosines in a different base and thirdly the amplification of these sequences which might be necessary for the measurement method assessed. All these steps influence the real expectation values and the results.

1. The production of methylated DNA may be incomplete and introduce errors. 2. The bisulfite conversion might be incomplete. This might introduce a crucial error, because unmethylated cytosines might appear as methylated cytosines if they were not converted properly. 3. The amplification in the PCR might be biased or have a high variance. All these steps add to any variance and/or bias in the final measurement method to be assessed and cannot easily be separated from it. In other words, these errors cannot easily be avoided.

The need in the art is obvious when it comes to interpretation of PCR based assay results, wherein the assays were designed to analyze methylation patterns and proportions of mehylated sequences in a mixture of methylated and unmethylated DNA sequences, such as in body fluid samples. Here it is a challenge to sensitively detect and quantify those sequences that are methylated in specific positions in presence of a high background of unmethylated sequences. After conversion, such as the bisulfite conversion, these sequences do differ in their base compositions at specific positions (in that they either comprise CpG or TpG dinucleotides). RealTime PCR assays use this difference by employing probes or primers or blockers that specifically bind to the one or the other dinucleotide. Because it was the common understanding that

diagnostically relevant CpG positions are found within a context of co-methylated CpG positions, i.e. a stretch of coordinated methylation occurring. Most assays are designed such that they rely on a number of several differentially methylated CpG positions in proximity, that are covered by specific blockers or primers and probes. In some situations however, this does not mirror reality. Samples that need to be analyzed may comprise of all kind of co-methylated or mosiac-methylated stretches. In the prior art however calibration of such measurement methods relied on standards that were "co-methylated" themselves. It is the subject of this invention to provide a calibration system that resembles the nature of a 'real sample' and takes into account that 'real samples' might not be co-methylated.

Assay calibration based on standards known in the art

created the problem, that a number of DNA molecules with aberrant sequences is not detected by these assays, that were designed to pick up methylated CpGs, even if only a part of these CpG dinucleotides is unmethylated and therefore converted into TpG dinucleotides. These however might still be of diagnostic relevance. The technical problem in the art is, that no methods or tools exist to address the extent of this potential lack of co-methylation, and it's effect e.g. on the diagnosis deducted from these results. The only way to address this problem was to design a plurality of assays covering either different CpG positions or being specific for different CpG/TpG occurrences and comparing the measurements. However to compare such different assays, a calibration standard would be needed, that simulated the situation in "real DNA", i.e. in a 'real sample'.

Only by providing a calibration standard as described in detail below, the test system allows to build models with patterns very close to observations in real DNA.

Therefore, one technical problem forming the basis of the present invention is to provide a calibration standard for determining base proportions of degenerated bases in DNA, and thereby enabling to assess at least two base proportions, preferably much more. In particular the technical problem is to provide a calibration standard for assessing measurement methods (for example assays, such as real time PCR based assays, or sequencing methods) determining specific base proportions in a mixture of DNA molecules.

The present invention solves this problem by providing a calibration standard for determining base proportions of degenerated bases in DNA, a degenerated base

representing at least two different bases in at least two DNA molecules at the same position, produced by a process comprising the steps of:

- providing at least two DNA molecules being not identical and containing at least two bases of the degenerated base at different positions within at least one DNA molecule; and
- mixing the DNA molecules in unequal ratios, thereby obtaining the calibration standard.

This calibration standard can now be used for assessing measurement methods, such as for example PCR based assays. Therefore it is one embodiment of the invention to provide a method for assessing a PCR based assay for its suitability when analyzing the methylation status of DNA.

Prior to setting forth the invention it may be helpful to an understanding thereof to set forth definitions of certain terms to be used hereinafter.

As used herein, the term "DNA" refers to a natural or synthetic polymer of single- or double-stranded DNA alternatively including synthetic, non-natural or modified nucleotides which can be incorporated in DNA polymers. Each nucleotide consists of a sugar moiety, a phosphate moiety, and a base moiety which is preferably either a purine or pyrimidine residue. In the present invention, the DNA is preferably of natural origin, such as genomic DNA and plasmids.

As used herein, the terms "DNA pool", "plasmid", "plasmid insert", "plasmid stock", "subclone", and "clone" are used interchangeably and refer to a homogeneous DNA of distinct length and sequence.

As used herein, "oligonucleotide" refers to a molecule comprising two or more deoxyribonucleotides or ribonucleotides, preferably more than three. The length of an oligonucleotide will depend on how it is to be used. Preferred is the range of 5 to 1000bp length. Also preferred is the range of 10 to 500 bp. More preferred is a range of 15 to 200 bp and most preferred is a range of 20 to 50 bp in length. The oligonucleotide may preferably be derived synthetically, however cloning is possible as well. Oligonucleotides may also comprise protein nucleic acids (PNAs).

Oligonucleotides can be synthesized using standard phosphoramidite chemistry. The degenerated bases are easily incorporated during synthesis. In addition, RNA oligonucleotides having more than approximately 30 nucleotides can be favorably synthesized in large amounts by in-vitro transcription. Synthesis and purification of oligonucleotides are well-known to those skilled in the art.

In the meaning of the invention, the phrases "degenerated base" or "mixed base" are used interchangeably and relate to at least two different bases in at least two DNA pools at the same position. Any of the at least two bases can be implemented into a single DNA molecule and thereby into a single DNA pool. The phenomenon of a mixed base is observed by regarding the DNA as a whole consisting of single DNA molecules. At least two bases are selected among bases composed of N-heteroaromates, preferably pyrimidine and purine bases, more preferably adenine (A), guanine (G), cytosine (C), thymine (T), uracil (U), and inosine (I), and modifications thereof. For example, degenerated bases are S (for C or G), W (for A or T), R (for A or G), Y (for C or T), M (for A or C), K (for G or T), H (for A, C, or T), B (for C, G, or T), D (for A, G, or T), V (for A, C, or G), and N (for A, C, G, or T). In the meaning of the invention, a degenerated base can also be a mixture of a natural base and a modification thereof, such as cytosine (C) and 5-methylcytosine (^{m}C).

As used herein, the phrase "base proportion" denotes a specific composition of DNA involving a plurality of base moieties at a specific position. The composition can take arbitrary values within a range from 0% to 100%. It is always related to the maximum possible occurrence of a single base moiety. That means, the exclusive occurrence of single base moiety along with the absence of other base moieties gives a base proportion of 100% related to the implemented base or 0% related to the non-implemented base, respectively. The base proportion can be determined by measurement methods which are able to quantify base compositions in DNA. Preferably, the single base moieties are detected separately. The proportion that is the basis for mixing these different DNA pools may then be calculated by a mathematical algorithm.

Surprisingly, it has been found that the standard according to the present invention is able to exhibit at least two base proportions at different positions. Furthermore, both proportions or even more can be advantageously set within a single calibration standard. The provision of the inventive calibration standard enables the measuring of different base proportions within one standard, which number can exceed both the number of possible bases or states (i.e. 2 = C or T in case of methylation analysis, or 4=A, C, T, G in case of mutation analysis) of the degenerated base and the number of DNA pools provided. The number of proportions which are theoretically possible is limited by the number of bases (states) of the degenerated bases to the power of provided DNA pools. The limiting value includes the terminal proportions of 0% and 100%, respectively, in which the bases of the degenerated base are identical in each DNA pool at the same position. The number of these terminal proportions corresponds to the number of bases (states) of the degenerated base. The terminal proportions can also be implemented into the inventive calibration standard and therefore be determined in the measurement method.

The number of proportions to be finally determined may preferably be controlled by the number of degenerated bases. The intention to record a maximum number of proportions within a single calibration standard requires at least an equal number of degenerated bases (i.e. positions of degenerated bases). For example if 4 different proportions are to be assessed at least 4 degenerated positions (degenerated bases) in the sequence are required. Additional degenerated bases do not further increase the number of measurable proportions, but may enhance the probability to record each possible proportion in case of their unequal distribution over the calibration standard.

The intensity and the level of base proportions can be advantageously altered via any unequal mixing ratio.

Therefore, the inventive concept of setting up numerous base proportions of a degenerated base over the entire range of a sequence that is addressed by an assay, or measurement method as calibration standard is of special benefit for assessing these measurement methods. It is also of special benefit for determining numerous base proportions of a degenerated base over the entire range of a measurement method to be assessed. The calibration standard is preferably applied in methods to quantify base compositions, such as sequencing, of DNA samples inherently

containing mixtures of bases at local positions, such as for example SNPs, or which are treated to contain mixtures of bases at local positions, such as cytosine and thymine (derived from 5-methylcytosine and unmethylated cytosine) after bisulfite treatment and amplification with a DNA polymerase, such as for example PCR. The determination of base proportions is of special interest to explain the biological function of molecule-to-molecule variations, such as differential methylation in real tissue samples. The accuracy of these methods can be reliably assessed and improved by the inventive standard.

In other words, the test system described here provides different proportions at different positions within one mixture. Therefore, it overcomes the problem of the other system of prior art, wherein equal proportions at all positions are used and thereby might bias measurements and not adequately mirror the situation in a real (i.e. naturally occurring) sample.

In addition, this inventive method allows to generate data over a range of measurement points and not only at one defined value, therefore a single mixture can be used to assess the whole range of a measurement method.

Such a range could comprise determining methylation values from 10%-90%. By providing one molecule mixture as calibration standard, that provides a number of different base proportions (ratios) at different positions (i.e. at different degenerated bases), for example in a methylation analysis assay : 1st CpG: 10%C and 90%T; 2nd CpG: 40%C and 60%T; 3rd CpG : 75%C and 25%T; 4th CpG 90%C and 10%T.

The at least two DNA pools underlying the inventive calibration standard can be either provided by oligonucleotide synthesis or generated from an inhomogeneous DNA, molecules of which are split and characterized to enable well-defined mixtures.

In the latter case, natural DNA can be prepared from samples by applying standard methods like lysis or heat treatment combined with phenol/chloroform extraction or purification by using silica based purification systems. Molecule-to-molecule variations of the inhomogeneous DNA, that are of interest to the measurement method to be assessed, are separated from each other. Favorably, a cloning procedure is applied to separate these variations and to allow generation of sufficient amounts of single specific DNA molecules for further processing. The sequences of the cloned DNA molecules are subsequently determined by appropriate methods, they may be aligned,

and a set of DNA molecules is chosen which are different at positions relevant for the measurement method to be assessed.

Contrary to prior art, the provided DNA molecules do not show the identical base at all degenerated base positions within one molecule but rather show patterns of different bases. They may show intramolecular differences concerning the occurrence of any base of the degenerated base at different positions within a single DNA molecule.

In a preferred embodiment the calibration standard to assess a measurement method suitable for a methylation analysis is generated by firstly selecting a number of different sources of natural occurring material expected to show different methylation patterns, i.e. from different tissues, organs or individuals and extract DNA from these. The extracted DNA molecules may be mixed and converted either by bisulfite treatment or incubation with the enzyme AID. If the effect of incomplete conversion is to be addressed in the final assay too, several tubes should be setup to be treated such that different conversion rates result. For example the treatment may be applied for different times and at different temperatures. Afterwards, the resulting mixture of a number of different sequences, differing only in their cytosine (C) versus thymine (T) content at positions which were unmethylated cytosines beforehand, is subjected to a cloning procedure. Thereby the single molecules are separated from each other and amplified (within their host). From these clones suitable amounts may be isolated for analysis. A detailed sequence analysis is performed to reveal which exact sequence is stored in which clone (or culture). DNA pools generated from these clones can then be mixed to specified ratios.

A number of two DNA pools is the minimum to be chosen, but three or even more DNA pools lead to a higher fidelity of assessment. The DNA pools are finally mixed in permutations of different ratios which allows to generate different mixtures from a constant number of DNA pools.

The method allows to generate test systems providing any wanted composition of base proportion at different DNA positions whenever a needed pattern can be found in subclones derived from real samples. This allows to always choose the appropriate subclones for any analysis method the test system will be applied to. It is e.g. possible

to choose stretches that show blocks with equal base proportions at all sites of interest. This way the influence of such blocks (like local co-methylation) on measurement methods can be assessed. Furthermore, any pattern can be simulated by designing the underlying base proportions using bioinformatics, synthesizing oligonucleotides and mixing them in pre-calculated ratios.

An established system according to the invention can easily be used as a standard for optimization and calibration experiments for different methods and is a potential commercial product. Once a test system like the one described is established it can easily and cheaply be reproduced with low effort and low risk of changes. More complex systems of prior art needing more preparation steps (than concentration measurement and mixing), e.g. random PCR or enzymatic preparation steps, are not as robust as the provided system according to the invention and have a high variance from batch to batch. All these characteristics make test systems based on the described method a potential commercial product: easy, reliably, and cheap to produce as soon as established.

In an embodiment of the present invention, the at least two DNA pools are provided by supplying inhomogeneous DNA and cloning it, thereby separating single molecules and amplifying these into homogenous DNA pools, determining the base composition differences of at least two DNA pools and selecting the at least two DNA pools.

Before downstream processing a suitable source of DNA is chosen which is known to be inhomogeneous with regard to the relevant base compositions. For example, the DNA can be obtained from body fluids of an individual. "Body fluid" herein refers to a mixture of macromolecules obtained from an organism. This includes, but is not limited to, blood, blood plasma, blood serum, urine, sputum, ejaculate, semen, tears, sweat, saliva, lymph fluid, bronchial lavage, pleural effusion, peritoneal fluid, meningal fluid, amniotic fluid, glandular fluid, fine needle aspirates, nipple aspirate fluid, spinal fluid, conjunctival fluid, vaginal fluid, duodenal juice, pancreatic juice, bile and cerebrospinal fluid. Experimentally separated fractions of all of the preceding are included as well. "Body fluid" also includes solutions or mixtures containing homogenized solid material, such as faeces. DNA can also be obtained from tissue sources, for example provided as clinical samples, such as tissue embedded in paraffin, histologic slides or fresh frozen tissue. These tissues may be for example, tissue from eyes, intestine, kidneys,

brain, heart, prostate, lungs, breast or liver, or all possible combinations thereof. Furthermore, DNA can be obtained by chemical synthesis.

Preferably, degenerated bases are an integral part of the DNA requiring the selection of identical regions with such local differences. Alternatively, sequence alterations may be introduced into purified homogenous DNA, thereby creating an inhomogeneous DNA which is to be considered to bear degenerated bases. Sequence alterations can be preferably obtained by chemical treatment or *in-vitro* mutagenesis.

For cloning, the DNA of interest is ligated into a suitable vector, preferably a plasmid. Before, it might be necessary to introduce restriction sites by amplification and/or ligation methods and/or to cut the DNA with restriction endonucleases. The vector is subsequently transformed into a suitable host, preferably a bacterial cell, such as E. coli. The cells are cultivated, preferably by spreading onto agar plates which results in subclones. A set of subclones is analyzed to obtain information about the base composition differences. In detail, the plasmid inserts are favorably sequenced to characterize the base implementation of degenerated bases in different DNA molecules at the same position. Cloning, transformation and sequencing procedures are well known to the skilled artisan. Finally, a minimum of two DNA molecules is selected for further generation of the inventive calibration standard.

In a preferred embodiment of the present invention, the inhomogeneous DNA is supplied by providing genomic DNA containing single nucleotide polymorphisms (SNPs).

Single nucleotide polymorphisms or SNPs are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. For example, a SNP might change the DNA sequence AAGGCTAA to ATGGCTAA. For a variation to be considered a SNP, it must occur in at least 1% of the population. SNPs which make up about 90% of all human genetic variation occur every 100 to 300 bases along the 3-billion-base human genome. Two of every three SNPs involve the replacement of cytosine with thymine. SNPs can occur in both coding (gene) and non-coding regions of the genome.

In the present invention, the DNA of different individuals is isolated and purified. The basic sequence of the genomic DNA samples is identical, only interrupted by local differences in terms of SNPs. Preferably, regions showing a higher frequency of SNPs are already known from the prior art, thereby allowing to focus on specified regions by previous amplification. The exact kind and positions of SNPs are determined in such a way described in the course of the specification. By mixing at least two DNA molecules of at least two individuals, the resulting calibration standards may be regarded as a population with well-defined and enriched SNPs. Viewing this population as a whole, the SNPs of individual DNA molecules can be regarded as the bases that are the degenerated bases in the meaning of the invention.

In another preferred embodiment of the present invention, the inhomogeneous DNA is supplied by chemically synthesizing an oligonucleotide containing degenerated bases.

In a preferred embodiment of the present invention, the degenerated base represents two bases in at least two DNA molecules at the same position. In a more preferred embodiment of the invention, the two bases are cytosine and thymine.

In another preferred embodiment of the present invention, the inhomogeneous DNA is provided by chemically treating DNA containing unmethylated cytosine bases in such a way that the unmethylated cytosine bases are converted to uracil, whereby said conversion may be incompletely performed and statistically evenly distributed. It is preferred that, and the obtained modified DNA is amplified afterwards. The chemical treatment is preferably conducted with a bisulfite (= disulfite, hydrogen sulfite).

The provided inhomogeneous DNA can be exclusively used for the generation of a calibration standard for determining base proportions of the degenerated base Y. The assessment of methods determining the cytosine/thymine base proportion is of special practical interest in quantitative DNA methylation analysis. In order to transfer methylation differences at cytosines within extracted DNA to amplifyable and detectable base differences appropriate methods have to be used. This is preferably done by a chemical treatment with bisulfite which will convert unmethylated cytosines to uracil, while methylated cytosines are unaffected. Thereby, the base composition of DNA is changed: Cytosines except the ones which were formerly methylated and which

are typically found in the sequence context CpG will be converted to uracil which may be replaced by thymine during amplification.

To provide a suitable DNA for the generation of a calibration standard for determining a plurality of base proportions of the degenerated base Y, depending on the intended use either the aforementioned treatment is incompletely performed, i.e. unmethylated cytosines are either converted to uracil or remain cytosine or DNA that is methylated to different degrees (i.e. that provides different proportions of methylation at specific sequence positions) at different CpG positions, is obtained and mixed before treatment. It is also possible to model the required pattern, because positions which are accessible for enzymatic methylation can be methylated by suited bacterial strains in-vivo or methylases in-vitro. That way specific patterns necessary for the calibration of certain assays may be generated.

To influence the conversion rate the treating conditions may be varied by several reaction parameters, such as period of incubation, concentration of treating agent, temperature, etc. Usually, an equal distribution of deaminated cytosine (= uracil) and original cytosine is expected intermolecularly and intramolecularly. Deviations form the equal distribution may occur, for example due to the specific sequence environment of a certain cytosine, and are tolerated if an identical intramolecular conversion is to be excluded. The converted DNA containing uracil is favorably converted a second time into common DNA containing thymine instead of uracil, due to a DNA polymerase based amplification process whereby uracil is read as thymine and therefore 'coupled' with adenine in the first round of replication. The conversion can be performed by PCR which simultaneously amplifies the treated DNA. However, this step is dispensable since plasmids carrying inserts of treated DNA are copied during the cloning process, thereby converting uracil to thymine as well. The inventive pattern of uracil (or thymine, respectively) and cytosine forms the basis for numerous C/T proportions at the same and/or different positions by the inventive mixing. The pattern is also required for mixing more than two DNA molecules in case of two bases of the degenerated base only, for mixing more than three DNA molecules in case of three bases of the degenerated base, etc. The fact that real sample material can be used for the initial step of subclone generation allows to easily reproduce patterns as observed in nature. E.g. for methylation analysis this offers the opportunity to test sensitive detection methods very precisely and in detail, and allows modeling reality in a more appropriate way than by mixing DNA of 0% and 100% methylation at all positions.

The initial DNA to be treated can be either methylated or unmethylated. If unmethylated DNA is used, this may be synthesized by chemical methods as already described in the course of the specification or may be derived from a genomewide DNA amplification method, as described in patent application DE 04 090 037 (and in PCT/EP2005/001407).

In a preferred embodiment of the present invention, the inhomogeneous DNA is amplified. The DNA, modifications thereof, or fragments thereof are amplified causing an increase in the number of copies of a particular DNA of interest and resulting in a particular DNA of interest which is of distinct length and consistently double-stranded.

The inhomogeneous DNA sample is purified to remove disturbing substances, such as inhibitors of the DNA polymerase or inhibitors of hydrogen bond formation, or substances promoting the formation of secondary and tertiary structures. Such downstream-processing is preferably performed by the method of precipitation, dialysis, gel filtration, gel elution, or chromatography, such as HPLC or ion exchange chromatography. It is recommended to combine several methods for better yields.

Preferably, amplification is performed by polymerase chain reaction (PCR). It is further preferred that at least 20 PCR trails are performed if using DNA of natural origin, whereby different regions of the same DNA are amplified. The resulting diversity of amplification products enhances the choice of subclones.

It is a particularly preferred embodiment of the invention that all steps of the amplification procedure take place in one tube, or well, or other suitable vial. All necessary components may be added at once comprising buffer, dNTPs, primer, a DNA polymerase and the DNA sample to be amplified. The chance for contaminations during the amplification procedure is eliminated to the greatest possible extent.

As used herein, the phrase "DNA polymerase" refers to enzymes that are capable of incorporating nucleotides onto the 3' hydroxyl terminus of a nucleic acid in a 5' to 3' direction thereby synthesizing a nucleic acid sequence. Examples of DNA polymerases that can be used comprise E. coli DNA polymerase I, the large proteolytic fragment of E. coli DNA polymerase I, commonly known as "Klenow" polymerase, "Taq" polymerase, T7 polymerase, Bst DNA polymerase, T4 polymerase, T5 polymerase,

reverse transcriptase, exo-BCA polymerase, etc. In the scope of the invention it is also possible that an inhomogeneous RNA is originally present, such as an oligoribonucleotide in particular, which is reverse transcribed into cDNA and may be further amplified, favorably in PCR.

In embodiments, wherein nucleic acids are amplified, it may be desirable to separate the amplification product from the template and the excess primer for the purpose of determining whether specific amplification has occurred. In one embodiment, amplification products are separated by agarose, agarose-acrylamide, or polyacrylamide gel electrophoresis using standard methods (Sambrook et al., In: Molecular Cloning: A Laboratory Manual 2nd rev. ed., Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 1989). Alternatively, chromatographic techniques may be employed to effect separation. There are many kinds of chromatography which may be used in the present invention: adsorption, partition, ion-exchange and molecular sieve, and many specialized techniques for using them including column chromatography.

Amplification products may be visualized in order to confirm amplification of the sequence of interest. One typical visualization method involves staining of a gel with ethidium bromide and visualization under UV light. Alternatively, if the amplification products are integrally labeled with radio- or fluorometrically-labeled nucleotides, the amplification products can then be exposed to x-ray film or visualized under the appropriate stimulating spectra, following separation. Advantageous radioactive isotopes are ^{3}H, ^{14}C, ^{32}P, ^{33}P, ^{35}S, or ^{125}I. Fluorescence dyes are well-known in the art.

In another embodiment, visualization is achieved indirectly. Following separation of amplification products, a labeled nucleic acid probe is brought into contact with the amplified sequence. The probe preferably is conjugated to a chromophore, but may be radio-labeled as well. In another embodiment, the probe is conjugated to a binding partner, such as an antibody or a low molecular weight ligand, and the other member of the binding pair carries a detectable moiety. Advantageous low molecular weight ligands for labeling nucleic acids are steroids, such as digoxigenin, biotin, and derivatives thereof. Digoxiginin (DIG) is a steroid hapten which does not occur in nature. Therefore, undesired side reactions are avoided by using the monoclonal anti-DIG antibody which is additionally characterized by a high sensitivity and specificity. D-biotin is bound with a remarkably high affinity of 10^{-15} M by streptavidin from

Streptomyces avidinii which is a homo-tetrameric protein containing a biotin binding site in each subunit. Both, the monoclonal anti-DIG antibody and streptavidin can be conjugated with reporter enzymes selected from the group consisting of peroxidase, CAT, GFP, GST, luciferase, β-galactosidase, and alkaline phosphatase. The antibody conjugated to the probe is preferably recognized by a second antibody conjugated with one of the aforementioned reporter enzymes.

In one embodiment, detection is performed by southern blotting and hybridization with a labeled probe. The techniques involved in southern blotting are well known to those of skill in the art. Briefly, amplification products are separated by gel electrophoresis. The gel is contacted with a membrane, such as nitrocellulose, permitting the transfer of the nucleic acid and non-covalent binding. Subsequently, the membrane is incubated with a chromophore-conjugated probe that is capable of hybridizing with a amplification product of interest. Detection is performed by exposure of the membrane to x-ray film or ion-emitting detection devices.

Following cloning, base composition differences are determined by sequencing in another embodiment of the present invention. It represents a convenient and reliable analysis for gaining sequence information. Appropriate methods, such as the deoxy method, are well-known to those skilled in the art. Preferably, 20 subclones of a single PCR are analyzed.

According to the invention the different variations of sequences are now characterized to as much detail as is necessary for the intended purpose. For example when analyzing SNPs, i.e ratios of bases at a specific position, the calibration standard for the according assay may only be analyzed to the extent that is required for the SNP analysis measurement method, which might be as little as the sequence analysis of one primer length, which may be extended by one base, i.e resulting jn a sufficient analysis of about 30bp. However, when analyzing the methylation pattern of a CpG island a larger region of sequence needs to be analyzed. For example, a so called HeavyMethyl-MethyLight assay may be used, which may analyze from 5 up to 20 different CpG positions within one assay (Cottrell et al., A real-time PCR assay for DNA-methylation using methylation-specific blockers. Nucleic Acids Res. 2004 Jan 13;32(1):e10). In addition there will be a number of cytosine positions that may be

converted sufficiently, or may not and may therefore appear as C or T. For these assays the region of interest may be 80 – 200 bp in length.
In another example the measurement method to be assessed may be the sequencing analysis of molecule mixtures. An adequate calibration standard for this measurement method should have been characterized in detail in an area of up to 500 bp or more.

In addition to the previous embodiments of supplying inhomogeneous DNA and processing it, in order to achieve at least two clones of pure homogenous DNA according to the invention at least two DNA pools can also be directly provided. Therefore, in another embodiment of the present invention, the at least two DNA pools are provided by chemically synthesizing different oligonucleotides in separated synthesis.

By mixing either the clones derived from initially inhomogenous DNA or the synthetically generated DNA molecules in unequal ratios, the phenomenon of degenerated bases is created with regard to the resulting calibration standard. In accordance with the presetting of the measurement method to be assessed, such as the metering range and the real sample, any pattern of degenerated bases can be purposively obtained. The pattern is characterized by the kind and arrangement of degenerated bases, and the proportions of bases of the degenerated base.

Preferably, the oligonucleotides of different synthesis contain homologous stretches of any length which are separated by non-identical stretches of any length. After mixing, the pattern of the degenerated base mainly depends on the non-identical stretches. Homologous stretches contribute to the diversity of arrangements of degenerated bases and the stability of the oligonucleotide. However, they are dispensable as the case arises. Preferably, a single kind of degenerated base in the calibration standard is desired. In addition, two bases underlying the base proportion of the degenerated base are preferred. For example, quantitative information concerning methylation is represented by cytosine/thymine proportions (after bisulfite conversion of unmethylated cytosine to uracil and PCR). Therefore, a standard consisting of DNA molecules exclusively composed of cytosine and thymine might be sufficient. Nevertheless, other bases should be incorporated as homologous stretches in order to increase the stability of depurinated oligonucleotides.

Depending on the intended use of such a calibration standard, the standard may also be designed to mimic a real sample, i.e. naturally occurring and treated for methylation analysis, to an extent as high as possible.

Depending on the length of the oligonucleotides, the yield of synthesis may require amplification and cloning of the oligonucleotides. Additionally, the handling of oligonucleotides as plasmid inserts is more convenient.

The intensity and level of base proportions within the oligonucleotide-base standard are further determined by the mixing ratio. To generate the greatest variety of sequences and ratios of specific base pairs, it is preferred that said mixing ratio is pre-calculated.

The provision of at least two synthesized oligonucleotides has a couple of advantages: The exact setting of any favored pattern prevents the doubling of patterns and allows to cover the complete measurement range Therefore, short DNA molecules can be provided for calibration of measurement methods for short molecules. In contrast to the supply, splitting and characterization of inhomogeneous DNA, oligonucleotides are easy and cheap to produce.

In a preferred embodiment of the invention, at least three DNA molecules are provided. As already stated, the maximum number of proportions follows a potential relation wherein the number of provided DNA molecules represents the power. The provision of three DNA molecules cubes the number of bases of the degenerated base. Assuming the preferred embodiment of the degenerated base which is only represented by two bases at the same position of different DNA molecules, eight proportions are distributed over the metering range by providing three molecules (2^3). Contrary, only four proportions are realized by providing two molecules (2^2). Furthermore, the provision of at least three DNA molecules enables a higher number of permutations, being a tool for maximizing base proportion numbers. Actually, the maximum number of proportions can be observed if the corresponding number of degenerated bases is available. However, the rectangular distribution of base proportions requires some more sampling in terms of degenerated bases. Likewise, an unequal distribution by using natural sources for DNA molecule provision makes actions necessary to cover as many measurement points as possible with high probability. In addition to an enhancement of the number of degenerated bases, further base proportions can be

advantageously gained from the same source by permuting the DNA molecules in the mixture. The number of permutations follows the factorial function. It is limited to 2 by two DNA molecules, but already increased to 6 by three DNA molecules.

In another preferred embodiment of the present invention, the DNA molecules contain at least 40 degenerated bases, preferably at least 145 degenerated bases. It is left to chance which base of a degenerated base is finally implemented in single DNA molecule. Therefore, the intramolecular pattern is variable and only predictable for a single DNA molecule with a certain probability. The total number of realizable patterns depends on the number of bases of the degenerated base and the number of degenerated bases. Favorable patterns have to be selected to provide DNA molecules according to the invention which are capable of forming a calibration standard with a maximum of base proportions after mixing. In order to diminish cloning and selection efforts for this purpose, the probability of providing the totality of possible base proportions for a special degenerated base and a given number of DNA molecules can be enhanced by increasing the number of degenerated bases. Favorably, the number of clones is equal to the number of needed DNA molecules. In case of two bases of the degenerated base and the provision of two DNA molecules taken from only two clones, a minimum number of 40 degenerated bases is required to guarantee all possible base proportions. This number is further increased to preferably 145 degenerated bases if providing three DNA molecules from three clones.

In one embodiment of the invention, DNA pools are provided showing an identity of less than 60% at the positions having implemented bases of the degenerated base, preferably less than 23%. These values of identity are related to the non-identical stretches consisting of degenerated bases to be measured, thereby excluding stretches betwixt regardless of which homology. The values serve as orientation for DNA molecule selection. In a preferred embodiment of the present invention, DNA molecules are provided showing the lowest possible identity at the positions having implemented bases of the degenerated base.

In another preferred embodiment of the invention, DNA molecules are provided containing each base of the degenerated base at different positions within each DNA molecule. Therefore, DNA molecules intramolecularly lacking at least a single base of the degenerated bases are excluded. It is an essential precondition to achieve the

maximum number of base proportions which is calculated by the number of bases of the degenerated bases to the power of provided DNA molecules.

In another preferred embodiment of the invention, the DNA pools are mixed in a ratio based on a numerical series of b^n, b being the number of bases or states of the degenerated base, and n being the set of nonnegative integers from 0 to the difference of the number of provided DNA pools and 1. Before mixing, DNA molecule stocks are adjusted to identical concentrations. Herein, the mixing ratio is another highly favorable parameter to set the number of base proportions to be measured. As result of the inventive mixing, the maximum number of possible base proportions can finally be measured. Additional precondition are the corresponding number of degenerated bases and the intramolecular variability concerning the implemented bases of degenerated bases. The three parameters are interlinked influencing among each other. The DNA molecules have to be mixed at least in the aforementioned ratio. Higher ratios do not further increase the number of base proportions, but may decrease it. An equal distribution of measuring points over the metering range is achieved by preferably mixing according to the above ratio. Preferably, the degenerated base represents two bases so that the provided DNA molecules are mixed in ratios of 1 : 2 : 4 : 8 etc. More preferably, three DNA molecules are mixed in ratios of 1 : 2 : 4 resulting in a mixture of 7 parts and at most eight base proportions from 0% to 100% in steps of 1/7, i.e. 0% (0/7), 14.3% (1/7), 28.6% (2/7), 42.9% (3/7), 57.1% (4/7), 71.4% (5/7), 85.7% (6/7), and 100% (7/7). Other ratios of at least 1 : 2 : 4 : 8 etc., such as 1 : 3 : 5 : 9 etc., are applicable as well. If the degenerated base represents three bases; the provided DNA molecules are preferably mixed in ratios of 1 : 3 : 9 : 27 etc.

Object of the invention is also a kit for determining base proportions of degenerated bases in DNA comprising a calibration standard according to the present invention and optionally, instructions for use of the kit. Preferably, the kit contains the calibration standard ready-for-use, for example in a suitable concentration. Depending on the method to be assessed and/or the steps of the method to be assessed, the standard consists of different base variations. When analyzing methylation it will consist of C/T variations either at CpG positions only, or at all C positions. In one preferred embodiment of the present invention, the kit comprises a buffer to carry the calibration standard for shipping and/or measuring.

It is another object of the present invention to use the calibration standard according to the invention for determining base proportions of degenerated bases in DNA. The measured data obtained from the calibration standard are compared to the expected values based on the generated base proportions. In contrast to prior art, the here provided test system allows to assess measurement methods as a whole or their single steps. It therefore provides detailed information about single steps and can locate error sources more easily than methods that provide only an assessment of a whole pipeline of steps. Regarding the analysis of methylation patterns for example, the corresponding measurement method can be subdivided into the single steps of bisulfite treatment and detection. Preferably, only the detection system is assessed. Alternatively, the calibration standard can be initially introduced into the method, thereby running through all steps.

The invention also relates to a method for determining base proportions of degenerated bases in DNA, a degenerated base representing at least two different bases in at least two DNA molecules at the same position, comprising the steps of:

- providing trails each containing the DNA, a DNA polymerase, a sequencing primer with a label corresponding to any base moiety, 2'-monodeoxy-NTPs, and a 2',3'-dideoxyanalog, whereby the 2'-monodeoxy-NTPs are contained in excess compared to the 2',3'-dideoxyanalog;
- DNA-dependent extension of the sequencing primer by a DNA polymerase, whereby fragments of different length with the dideoxyanalogs at the 3' terminus are obtained;
- unifying the trails;
- separating the fragments; and
- detecting the labels, thereby determining the base proportions of degenerated bases,

wherein the calibration standard according to the present invention is used.

The determination of base compositions is favorably performed by sequence analysis according to the deoxy method of Sanger et al. (1977) PNAS USA 74, 5463-5467. It is based on the controlled interruption of enzymatic replication by 2',3'-dideoxyanalogs which incorporation blocks up further strand growth. A truncation at each position in

different molecules is caused by optimized reaction conditions resulting in a pool of truncated fragments which differs in length among each other by a single nucleotide. The fragments are separated by length and detected. Initially, the method is performed in a couple of trials depending on the number of base types underlying the sequence. Each trail contains the DNA to be analyzed, a DNA polymerase, a sequencing primer with a label corresponding to a certain base, 2'-monodeoxy-NTPs, and a 2',3'-dideoxyanalog which is preferably selected from the group of 2',3'-didesoxy-ATP, 2',3'-didesoxy-GTP, 2',3'-didesoxy-CTP, and 2',3'-didesoxy-TTP. Compared to the concentration of the 2',3'-dideoxyanalog, the 2'-monodeoxy-NTPs are supplied in excess for an equal statistical truncation during replication. The sequencing primer is extended template-dependently by means of a DNA polymerase according to the present specification. The fragments are characterized by their different length, the labeling at the 5'-end and the dideoxyanalog at the 3'-end, whereby the latter features are specific for each trail and each base type. The trails are merged and the fragments separated, favorably by loading on a denaturing polyacrylamide gel. The shorter the fragment, the faster the time of retention. By passing the detector position, the labels are recognized and assigned to a specific base type. Preferably, the sequencing primers are labeled with fluorescence dyes emitting different wave length. Fluorescence stimulation is caused by the absorption of energy, preferably provided by radiation, which is released again as photon with a shift in wave length of 30 to 50 nm, and within a period of approximately 10^{-8} seconds. The color sequence corresponds directly to the base sequence of the complementary strand in 5' to 3' direction.

Another object of the invention is a method for production of a calibration standard for determining base proportions of degenerated bases in DNA, a degenerated base representing at least two different bases in at least two DNA molecules at the same position, comprising the steps of:

- providing at least two DNA molecules being not identical and containing at least two bases of the degenerated base at different positions within at least one DNA molecule; and
- mixing the DNA molecules in unequal ratios, thereby obtaining the calibration standard.

It is another object of the present invention to provide a method for calibration of measurement systems which determine base proportions of degenerated bases in DNA, wherein a degenerated base represents at least two different bases in at least two DNA molecules at the same position, characterized by the use of a calibration standard according to claims 1-16 comprising the steps of

- performing said measurement system with one or several of said calibration standards at least once,
- determining a calibration curve or function
- performing said measurement system with the sample to be analyzed
- comparing the result with those of step 1 and step 2 and
- assessing the measurement system performed.

A preferred embodiment is a method according to claim 20 for calibration of measurement systems which determine the proportions of cytosine and thymine at positions, which show a degenerated base following conversion of unmethylated cytosines, characterized by the use of a calibration standard according to claims 7-16.

All embodiments of the calibration standard have been characterized in detail in the course of the present specification. The prior teaching of the present invention and embodiments thereof are considered as valid and applicable without restrictions to the method for production of the calibration standard if expedient.

The following figures 1 to 3 illustrate the present invention of a method to assess measurement methods quantifying base compositions in DNA.

Figure 1: Three final clones chosen for the mixtures, only genomic C positions and their on bisulfite treatment based equivalent (T) are shown.

Figure 2a: Number of measuring points for different C/(C+T) ratios within all six subclone mixtures of the example.

Figure 2b: Real calibration data based on an assessment of base ratio detection with four dye capillary sequencing.

Figure 3: Appendix: data from 96 subclones. Full sequence of clones from the initial sub-cloning step of G6e (part 1 to 3).

The following example is provided by way of illustration and not by way of limitation. Within the example, standard reagents and buffers that are free from contaminating activities (whenever practical) are used.

The method is hereby explained with help of an example experiment used for calibration. In the example experiment the test system was used to assess cytosine/thymine base ratio measurement methods as used in most methylation detection protocols using bisulfite treatment of the DNA.

First identical regions with local differences are subcloned into plasmids. In the experiment this was an inhomogeneous PCR product from incomplete bisulfite treated DNA that resulted in a mixture of molecules with different C/T proportions at all positions that were cytosine prior to conversion with bisulfite.

A set of the subclones is sequenced to obtain information about the base composition differences. Other methods to determine these differences can be used but sequencing of the subclones is the most appropriate method. In the experiment we sequenced 96 subclones from one inhomogeneous amplificate (Fig. 3).

A set of subclones is chosen that compared to each other are different at as many positions as possible relevant for the measurement method to be assessed. For this method a number of two chosen subclones is the minimum, but three or more lead to a higher resolution. In the experiment we chose three clones which differed at positions that in the genomic sequence were cytosine and resulted in either cytosine or thymine dependent on the bisulfite conversion (see Fig. 1).

Mixable amounts of the chosen plasmids are gained by cultivation of the subclones and plasmid preparations. The gained plasmid stocks are equilibrated to equal concentrations before mixing.

The plasmid stocks are mixed in unequal proportions. To gain more test mixtures from the same source the proportions are permuted. Though this is possible with many

proportions we suggest to use proportions based on 2^n; $n \in [1, 2 \ldots (\text{cloneNumber} - 1)]$. In the experiment we mixed the clones in the proportions 1 : 2 : 4, which resulted in eight equally distributed base compositions from 0/7 to 7/7 in steps of 1/7. Permuting the proportions allowed to generate six different mixtures (proportion permutations for six different mixtures of three clones: (1:2:4), (2:1:4), (1:4:2), (2:4:1), (4:1:2), (4:2:1)) from the three clones which in this experiment covered many measurement points at different levels (see Fig. 2a). A choice of four clones might be used for up to 24 mixtures with permutations of the proportions 1 : 2 : 4 : 8 leading to 16 base compositions from 0 to 1 in 1/15 steps.

The mixtures (in the experiment six) can now be used to assess or calibrate methods which measure the base proportions at specific positions. Results from the method to be calibrated or assessed can be compared with expectation values based on known proportions in the test system. An example for this is given in Fig. 2b.